THE WEB SCRAPER'S WORLD OF COPYRIGHT EXCEPTIONS AND CONTRACTUAL OVERRIDES

DARREN **ANG***

"Web scrapers" are software programs designed to automate the copying of data from the Web, and they can be used for a variety of useful applications. By building upon a technical understanding of how web scrapers operate, this article discusses the two main areas of law which control the use of web scrapers ("web scraping"): copyright law, which provides for certain exceptions that may be relied upon by users of web scrapers to avoid issues of copyright infringement, and contract law, which might bind users of web scrapers under webpages' terms of service and restrict them from conducting web scraping activities. Following which, this article turns to the growing number of statutory regimes which prohibit the use of contractual terms to exclude the operation of copyright exceptions, and considers the circumstances under which web scrapers may rely on these regimes.

I. INTRODUCTION

"Web scrapers",[1] technically defined, are software programs which are "designed to automate the downloading and parsing of the data from the Web".[2] They operate by interacting with the Web's communication protocols to extract data from webpages,[3] and they can do so at astonishing speeds. The data extracted from the use of web scrapers ("web scraping") can then be used for a wide variety of applications.[4]

---

* LL. B. (Hons) Candidate at the National University of Singapore; LL. M. Candidate at the University of Melbourne. This article would not have been possible without a proper appreciation of the technical details involved in web scraping—in that regard, I extend my most sincere gratitude to the Python programming communities on YouTube and StackOverflow for creating such wonderful programming tutorials, and making them easily and freely available. I also wish to extend the most heartfelt gratitude to Jerome Tan for the wonderful editing work and helpful comments on this piece. All errors remain my own.

[1] It is notable that "web scrapers" have been referred to in a variety of ways in the case law, such as "scraping programs", "screen scrap[ers]", "webcrawlers", or "spider[s]"—see Andrew Sellars, "Twenty Years of Web Scraping and the Computer Fraud and Abuse Act" (2018) 24:2 BU J Sci & Tech L 372 at 381-382.

[2] Jay M Patel, *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*, (California: Aspress, 2020) at xvii. For the avoidance of doubt, this article shall refer to web scraping algorithms (the bots themselves) as "web scrapers", the ones who program or code those algorithms as "programmers of web scrapers", and the human users of web scrapers (such as an app developer who uses a web scraper's output to create an application) as "users of web scrapers".

[3] A technical explanation of this process is provided in Part II of this article. Note that the term "webpage" has been preferred over "website"; from a technical perspective, a "webpage" refers to a single block of code that may be rendered on a web browser, while a "website" usually refers to a collection of webpages which are connected in some way—see "What is the difference between webpage, website, web server, and search engine?" (8 October 2021), *MDN Web Docs* (blog), online: <https://developer.mozilla.org/en-US/docs/Learn/Common_questions/Pages_sites_servers_and_search_engines>.

[4] Sellars, *supra* note 1 at 374.

However, some of these applications are more desirable than others. On one hand, some of the most ground-breaking features of the Web, such as search engines and optimised search results, require the use of web scrapers for their implementation.[5] On the other, web scrapers have facilitated fraudulent acts such as the creation of fake accounts for credit card scams.[6] Additionally, a carelessly-programmed web scraper might cause a website to be overloaded with indiscriminate traffic, causing it to slow down or go offline entirely.[7] These "distributed denial-of-service attacks", or as they are more commonly known, "DDoS attacks", attract criminal liability in most jurisdictions.[8] While it is a standard practice for programmers of web scrapers to set a reasonable time delay between requests to the Web,[9] the risks to website owners still remain.

Therefore, for societies to reap the benefits of web scrapers without bearing too much of their risks, entire arsenals of legal controls have been deployed against the use of web scrapers, as well as certain applications created upon web scrapers' outputs.[10] This article shall discuss the current state of the law regarding two of those controls: copyright law and contract law. Its primary aim is to provide readers with a general map of the complex frameworks covering this area, in a manner that may accommodate the regimes of as many jurisdictions as possible. It does so by drawing examples from multiple jurisdictions and attempting to define common threads.

Following this introduction, Part II sets the foundation of this article by providing a technical overview of how web scrapers can be used to create useful applications. Parts III and IV build

---

[5] For an excellent technical summary of some real-world applications of web scrapers, see Patel, *supra* note 2 at 1-30.

[6] Christopher Watkins, "Web Scraping Fraud: Going, Going … Ongoing" (10 September 2019), *Datavisor* (blog), online: <https://medium.com/datavisor/web-scraping-fraud-going-going-ongoing-c0f7a0db7310>.

[7] Danny Palmer, "What is a DDoS attack? Everything you need to know about Distributed Denial-of-Service attacks and how to protect against them" (15 October 2020), *ZDnet* (blog), online: <https://www.zdnet.com/article/what-is-a-ddos-attack-everything-you-need-to-know-about-ddos-attacks-and-how-to-protect-against-them/>.

[8] See e.g. Singapore: *Computer Misuse Act* (Cap 50A, 2007 Rev Ed Sing), s 7; UK: *Computer Misuse Act 1990* (UK), c 18, s 3; US: 18 USC § 1030 (2018).

[9] See Patel, *supra* note 2 at 60.

[10] In addition to statutes which deal with DDoS attacks, one major area of contention concerning the use of web scrapers is in personal data protection—see e.g. "Facebook says hackers 'scraped' data of 533 million users in 2019 leak", *The Straits Times* (7 April 2021), online: <https://www.straitstimes.com/world/united-states/facebook-says-data-on-530-million-users-scraped-before-september-2019>. While it is an incredibly fascinating area of law to look into, it is, most unfortunately, beyond the scope of this article.

upon this technical foundation, and discusses the copyright and contractual considerations of web scraping respectively. Part V discusses the intersection between these two dimensions, found in the prohibitions against using contractual terms to override the operation of statutory copyright exceptions. Finally, Part VI summarises the implications of these legal frameworks for users of web scrapers, and concludes the paper with some final thoughts about the ethics of web scraping.

II. HOW WEB SCRAPING WORKS

At the outset, a distinction must be drawn between the use of web scrapers themselves and the creation of applications based on web scrapers' outputs.

In short, a web scraper merely creates copies of webpages' source codes, which are unlikely to be of much use on their own. However, the source codes may then be processed and analysed in ways that result in useful applications. These two steps often come hand-in-hand, and they shall be discussed in turn below.

A. Step 1: The use of web scrapers in themselves

1. How web scrapers work: A technical overview

We begin with web scrapers themselves. Conventionally,[11] most web scrapers are programmed to do a single task: they look up the underlying source codes of webpages and copy the parts of those webpages which are relevant to the programmer.

The source code which all webpages are built upon is known as HyperText Markup Language, or HTML.[12] When we surf the Web on a web browser, the "pages" which we see on our screens are formed through the rendering of webpages' HTML codes—that is, every word, image, or video

[11] The technical explanations in Part II of this article are based on the conventions used by Python programmers, as it is generally accepted that the Python programming language is the industry standard language for data science and machine learning—see e.g. "Why is the Data Science Industry Demanding Python?" (10 February 2020), *Institute of Data* (blog), online: <https://www.institutedata.com/blog/why-is-the-data-science-industry-demanding-python/>. However, similar processes should apply for other programming languages.

[12] Patel, *supra* note 2 at 31.

which appears on a webpage would have been derived from some expression in HTML code.[13] As a classic example, for a web browser to display a simple webpage titled "Hello", with the plain text "Hello, world" on it, the following HTML code would have to be rendered:

```
<html><title>Hello</title><body>Hello, world</body></html>
```

Webpages are accessed through their unique Uniform Resource Locators, or URLs.[14] By sending a "request" to a particular URL, a web browser may identify the HTML code of the webpage at that URL and render it. The same process is used by web scrapers to copy the HTML codes of webpages—they are designed to send "requests" to webpages and *copy* the HTML codes associated with them.[15] The programmer determines the parts of the HTML code which are relevant for the application they intend to create, and copy only those parts.[16]

2. *The Robots Exclusion Protocol and "robots.txt" files*

No discussion of web scraping is complete without a brief nod to the Robots Exclusion Protocol (the "REP").[17] Under this protocol, website owners may include a "robots.txt" file on their website,[18] which tells web scrapers which parts of it they are permitted to scrape based on certain conventions.[19]

---

[13] The reader may experience this by opening any webpage and finding the "View page source" setting for the browser in use—this setting shows the underlying HTML code of the webpage that is currently open. It should be noted that this article only focuses on the legal protections surrounding the HTML code; the words, images and videos that result from the rendering of HTML code might be protected by different copyrights or contractual terms.

[14] Note that the technical details regarding accessing webpages through web browsers are much more complicated than this brief description might suggest—webpages are located at a specified "Internet Protocol Address", which might be rendered in text form through the "Domain Name System". These are standard protocols that the Web is run upon.

[15] Patel, *supra* note 2 at 37.

[16] *Ibid* at 41-42.

[17] Google has expended significant efforts in making the REP into an Internet standard—see Henner Zeller, Lizzi Harvey & Gary, "Formalizing the Robots Exclusion Protocol Specification" (1 July 2019), *Google Search Central Blog* (blog), online: <https://developers.google.com/search/blog/2019/07/rep-id>.

[18] On the difference between "webpages" and "websites", see n 3 of this article.

[19] This is the author's simplification; for a slightly-more technical description, see Patel, *supra* note 2 at 59.

Compliance with the REP is strictly voluntary, and programmers of web scrapers are free to ignore websites' "robots.txt" files if they choose to do so.[20] However, as a matter of etiquette, the general practice is for programmers to ensure that their web scrapers comply with websites' "robots.txt" files;[21] major search engines such as Google and Bing are programmed in this manner.[22] Google estimates that there are about half a billion websites which rely on this protocol.[23]

B. Step 2: The use of web-scraped HTML code for useful applications

Once a web scraper has copied the HTML code of a webpage, this code may then be processed and used for a wide variety of applications. For example, Google's search engine is powered by the processing of huge amounts of webpages' HTML codes, which are obtained through web scrapers "run simultaneously by thousands of machines".[24] Large amounts of HTML code can also be used in conjunction with statistical techniques to yield empirical conclusions,[25] which in turn, can be used for a variety of purposes.

III. COPYRIGHT EXCEPTIONS RELEVANT TO WEB SCRAPING

A. Why web scrapers might infringe copyrights

The most immediate legal concern for users of web scrapers is that their web scraping activities might infringe the copyrights of webpage owners. Computer code, which includes HTML, is explicitly classified as a type of work that may be protected by copyright under the *TRIPS Agreement*, which currently has 164 state parties.[26] With this in mind, the ability of web scrapers (in

---

[20] "Can I block just bad robots?", *The Web Robots Pages (blog)*, online: <https://www.robotstxt.org/faq/blockjustbad.html>.

[21] Patel, *supra* note 2 at 387.

[22] Google: "Introduction to robots.txt", *Google Search Central Documentation*, online: <https://developers.google.com/search/docs/advanced/robots/intro>; Bing: "robots.txt tester", *Microsoft Bing Webmaster Tools help & how-to*, online: <https://www.bing.com/webmasters/help/robots-txt-tester-623520ca>.

[23] Zeller, Harvey & Gary, *supra* note 17.

[24] "Advanced: How Search Works" (22 November 2021), *Google Search Central*, online: <https://developers.google.com/search/docs/advanced/guidelines/how-search-works>.

[25] See e.g. Daniel Seng, "The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices" (2014) 18 Va JL & Tech 369.

[26] See *Marrakesh Agreement Establishing the World Trade Organization*, 15 April 1994, 1869 UNTS 154 Annex 1C at art 9(2) at art 10(1) (entered into force 1 January 1995) [*TRIPS Agreement*].

themselves) to copy voluminous amounts of HTML code from multiple webpages might immediately raise concerns of copyright infringement in many jurisdictions across the world.[27]

However, not all HTML code is protected by copyright, as copyright protection only extends to "expressions" and not to "ideas".[28] In many jurisdictions, the line between "ideas" and "expressions" is drawn by the concept of "originality", which in turn, is a fact-sensitive inquiry that looks towards the creative efforts expended by the author of a work.[29] Therefore, when considering whether web scraping might result in copyright infringement, the real question to be asked is generally whether the HTML code copied by web scrapers is "original" enough to warrant copyright protection.

This turns out to be an extremely difficult question to answer. While only an "extremely low" level of originality is required for a work to be protected by copyright,[30] there are many features of HTML code which make even this low threshold hard to cross: the code might be seen as "factual", rather than creative, in nature;[31] or it might be seen as being primarily dictated by functional purposes rather than authorial creativity.[32] Ultimately, it is difficult to determine whether web scraping raises copyright issues, as it is difficult to determine whether specific blocks of HTML code may be protected by copyright in the first place.

Nonetheless, the vexed issues surrounding web scraping and copyright infringement can be avoided by turning to the limitations and exceptions found in most copyright statutes across the world (henceforth referred to as the "copyright exceptions"). This article shall discuss two copyright exceptions which are particularly relevant for web scraping: those for text and data mining ("TDM"), as well as the more open-ended "fair use" and "fair dealing" exceptions.

---

[27] See e.g. the facts of *The Newspaper Licensing Agency Ltd v Meltwater Holding BV* [2011] EWCA Civ 890 [*Meltwater*].

[28] See *TRIPS Agreement*, *supra* note 26 at art 9(2). For a domestic example, see *Global Yellow Pages Ltd v Promedia Directories Pte Ltd* [2017] SGCA 28 at para 15 [*Global Yellow Pages*].

[29] See e.g. *Global Yellow Pages*, *supra* note 28 at para 24; *Feist Publications, Inc. v Rural Tel. Service Co.*, 49 US 340 at 348 [*Feist*]; ECJ *Infopaq International A/S v Danske Dagblades Forening*, C-5/08, [2009] ECR I-6624 at para 37.

[30] *Feist*, *supra* note 29 at 345; *Global Yellow Pages*, *supra* note 28 at para 27;

[31] *Feist*, *supra* note 29 at 347. Incidentally, this line of reasoning was suggested by respondents to the public consultations for the Singapore *Copyright Act*—see Ministry of Law Singapore & Intellectual Property Office of Singapore, *Singapore Copyright Review Report* (17 January 2019) at para 2.8.1.

[32] *Computer Associates Intern., Inc. v Altai, Inc.* 982 F (2d) 693 at 709-710 (2nd Cir 1992).

B. Text and data mining exceptions

1. The policy of the TDM exceptions

To alleviate the legal uncertainty associated with TDM activities, which may include certain applications of web-scraped HTML code, many jurisdictions have introduced specific copyright exceptions for them.[33]

TDM activities are generally defined as automated or computational *analytical* techniques used to extract *new information* from existing data.[34] Crucially, some form of analysis must be done—the underlying data cannot be copied without anything more. A standard example of an act which would *not* be exempted by the TDM exceptions is the use of an automated tool (such as a web scraper) to collate works into a database, without performing any additional analysis on the works.[35]

In general, the TDM exceptions prevent TDM activities from giving rise to liability for copyright infringement when they are done appropriately. Their shared underlying policy is that TDM activities are beneficial for the economy and society-at-large, and as a result, copyright law should not inhibit them.[36] Conversely, returning to the example from the preceding paragraph, the mere use of automated tools to *copy* data, without performing any additional analysis on it, would not be consistent with the exceptions' underlying policy.

2. Common features of TDM exceptions across jurisdictions

---

[33] EC, Policy Department for Citizens' Rights and Constitutional Affairs, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects* (February 2018) at 12-13 [*TDM – Legal Aspects*].

[34] UK: *Copyright, Designs and Patents Act 1988* (UK), c 48, s 29A [*UK CDPA*]; EU: Directives EC, *Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC*, [2019] OJ, L 130/92 at art 3 [*EU CDSM Directive*]; Singapore: *Copyright Act 2021* (No 22 of 2021, Sing), s 244 [*Singapore Copyright Act*].

[35] See Ministry of Law Singapore & Intellectual Property Office of Singapore, *Public Consultation on Proposed Changes to Singapore's Copyright Regime* (23 August 2016) at para 3.64.

[36] UK: Ian Hargreaves, *Digital Opportunity: A Review of Intellectual Property and Growth* (May 2011) at para 5.26; EU: *TDM – Legal Aspects*, *supra* note 33at 14-15; Singapore: *Singapore Copyright Review Report*, *supra* note 31 at para 2.8.4.

While the exact scopes of the TDM exceptions adopted across jurisdictions differ, they share some common features, and two of them shall be discussed below.

First, the TDM exceptions are purpose-based. That is, the exceptions can only be relied upon if the relevant TDM activities fall within the spectrum of permissible purposes, which are delineated by the relevant statutory provisions. On the narrower end of the range of TDM exceptions is the UK's TDM exception, which only allows copies of a work to be made "for the sole purpose of research for a non-commercial purpose".[37] On the other end, the TDM exception in Singapore's *Copyright Act* allows for copies of a work to be made for the purposes of "computational data analysis", which in turn, is defined broadly and non-exhaustively.[38]

Second, most of the TDM exceptions require the person undertaking TDM activities to have "lawful access" to the works which are being mined.[39] While the exact scope of the term "lawful access" has not been defined within these exceptions, a standard example given by many jurisdictions is having a paid subscription to access a database of works.[40]

3. How the TDM exceptions might apply to cases involving web scraping

How, then, do the TDM exceptions apply to web scraping? It bears recalling that web scraping is a common precursor to the analysis of that code,[41] but it does not constitute any form of analysis in and of itself—it only copies the HTML codes of webpages in some form.

Therefore, the act of web scraping without anything more would not satisfy the TDM exceptions, and it is only the use of web scrapers *for the purpose* of creating applications upon the web-scraped data which may qualify for the exceptions, subject to the relevant requirements being met.

---

[37] *UK CDPA*, *supra* note 34, s 29A.

[38] *Singapore Copyright Act*, *supra* note 34, ss 243-244. This exception also applies to preparatory work done for the purposes of computational data analysis.

[39] A notable exception is Japan. See Tatsuhiro Ueno, "The Flexible Copyright Exception for 'Non-Enjoyment' Purposes – Recent Amendment in Japan and Its Implication" (2021) 70:2 GRUR International 145 at 149.

[40] UK: Intellectual Property Office, *Exceptions to copyright: Research* (October 2014) at 8; Singapore: *Singapore Copyright Review Report*, *supra* note 31 at para 2.8.5.

[41] See Sellars, *supra* note 1 at 373-375.

C. Open-textured "fair use" and "fair dealing" exceptions

Users of web scrapers might have another string to their bows—the open-textured "fair use" and "fair dealing" exceptions, which are available in more than 40 countries around the world.[42] These, at least on their face,[43] fall on a spectrum of flexibility: at one end are the narrower "fair dealing" exceptions which exhaustively state the purposes which may qualify for them,[44] and at the other are the broader "fair use" exceptions which adopt non-exhaustive factors as guidelines for courts to determine whether the use was fair.[45]

Despite the stark differences in statutory language, the "fair use" and "fair dealing" exceptions share a common theme. They both recognise that copyright is ultimately a balancing act: it must incentivize private individuals to make creative works, while also preserving the public interest in the dissemination of creative works.[46] Following from which, some works which appear to be copyright-infringing might actually promote the underlying goals of copyright under their specific circumstances.[47] The inquiry as to whether a use was "fair" ultimately collapses into a case-by-case inquiry that takes into account all the relevant factors.[48]

These exceptions are generally wide enough to encompass a whole range of activities, and it is unsurprising that they have been raised in some cases involving allegations of copyright infringement caused by web scrapers.[49] Therefore, if a particular jurisdiction does not have a TDM

---

[42] Jonathan Band & Jonathan Gerafi, "The Fair Use/Fair Dealing Handbook" (March 2015), *infojustice.org*, online: <http://infojustice.org/wp-content/uploads/2015/03/fair-use-handbook-march-2015.pdf> at 3.

[43] There is considerable academic literature surrounding the proposition that the "fair dealing" and "fair use" exceptions are not as different as the language of their provisions might suggest—see e.g. Ariel Katz, "Debunking the Fair Use vs. Fair Dealing Myth: Have We Had Fair Use All Along?" in Shyamkrishna Balganesh, Ng-Loy Wee Loon & Sun Haochen, eds, *The Cambridge Handbook of Copyright Limitations and Exceptions* (Cambridge: Cambridge University Press, 2021) 111. However, a discussion on this issue would extend far beyond the scope of this article—it suffices to note that the issue exists.

[44] See e.g. UK: *UK CDPA*, *supra* note 34, ss 29, 30 and 30A; New Zealand: *Copyright Act 1994* (NZ), 1994/143, ss 42-43.

[45] See e.g. Singapore: *Singapore Copyright Act*, *supra* note 34, s 191; US: 17 USC § 107.

[46] See e.g. Australia: *IceTV Pty Ltd v Nine Network Australia Pty Ltd* [2009] HCA 14 at para 24-25; Canada: *Cinar Corp. v Robinson* 2013 SCC 73 at para 23, [2013] 3 SCR 1168; US: *Altai*, *supra* note 32 at 711.

[47] See e.g. Singapore: *Global Yellow Pages*, *supra* note 28 at paras 74-76; US: *Campbell v Acuff-Rose Music, Inc.*, 114 S Ct 1164 at paras 3,4 (1994).

[48] This is true even in jurisdictions which feature the narrower "fair dealing" exceptions in their copyright statutes—see e.g. UK: *Ashdown v Telegraph Group Ltd* [2001] EWCA Civ 1142 at para 70; New Zealand: *Media Works NZ Ltd v Sky Television Network Ltd* (2007) 74 IPR 205 at para 74.

[49] See e.g. UK: *Meltwater*, *supra* note 27; US: *Field v Google Inc.*, 412 F Supp (2d) 1106 (D. Nev. 2006).

exception, or if an application created upon web-scraped HTML code would not satisfy the requirements of the TDM exception, the "fair use" and "fair dealing" exceptions might still be relied upon to avoid liability for copyright infringement.

IV. CONTRACTUAL RESTRICTIONS TO WEB SCRAPING

Although users of web scrapers might escape liability for copyright infringement by relying on the exceptions discussed above, they might find themselves facing *contractual* liability under contracts entered into between them and website owners.

The most common way for this liability to arise is under websites' terms of service. Additionally, considering the centrality of the REP to web scraping, the possibility of "robots.txt" files being incorporated into websites' terms of service shall be briefly considered.

A. Websites' terms of service

Many websites contain a page of terms and conditions that purport to govern their use (the "terms of service"). These terms of service might contain specific restrictions against the use of web scrapers,[50] or more general restrictions against the unauthorised use of any content hosted on the website.[51]

Naturally, users of web scrapers might face contractual liability for breaching these terms—this might mean having their access restricted or their account terminated.[52] Where applications are

[50] See e.g. the clause at issue in *HiQ Labs, Inc. v Linkedin Corp*, 938 F (3d) 985 at 991, n 5 (9th Cir 2019) [*Linkedin*]: more than 95 million automated attempts to scrape data were blocked every day, based on the clause in the terms of service which provided that users were not allowed to "[s]crape or copy profiles and information of others through any means (including crawlers, browser plugins and add-ons, and any other technology or manual work)", "[u]se manual or automated software, devices, scripts robots, other means or processes to access, 'scrape,' 'crawl' or 'spider' the Services or any related data or information", or "[u]se bots or other automated methods to access the Services".

[51] See e.g. the clause at issue in *PropertyGuru Pte Ltd v 99 Pte Ltd* [2018] SGHC 52 at para 77 [*PropertyGuru*]: the terms of service state that the site's content can "only be used for your own and non-commercial use, and not for publication, distribution, transmission, retransmission, redistribution, broadcast, reproduction or circulation to someone else in the same company or organisation, and not for posting to other websites or forums, newsgroups, mailing lists, electronic bulletin boards, or Internet Relay Chats operated by other websites".

[52] See *Linkedin*, *supra* note 50.

created upon web-scraped data, the creators of those applications might also face tortious liability for inducing the users of their applications to breach websites' terms of service.[53]

However, the above analysis proceeds on the assumption that the users of web scrapers are bound by websites' terms of service in the first place. This assumption does not always hold true. In general, the enforceability of websites' terms of service remains a controversial issue,[54] and it has been argued that the question should be answered with reference to basic principles of contract law.[55] In this regard, it is notable that a two-part test has been developed in US case law to guide the inquiry of whether a website's terms of service are enforceable against a user:[56]

(1) "[W]hether the terms were 'reasonably communicated' to the user; and

(2) [W]hether the terms were accepted by the user."

Determining whether the user of a web scraper is bound by a website's terms of service is a complicated exercise. However, once this hurdle is crossed, the user of the web scraper would generally be bound by the terms stipulated in the website's terms of service, which might include restrictions on their web scraping activities.

B. Websites' "robots.txt" files?

Unlike websites' terms of service, which have been found to have binding force in several cases, the legal status of "robots.txt" files remains entirely uncertain.[57] Nonetheless, as the REP is a highly-established industry standard for programmers of web scrapers, it is conceivable that a case may arise in future where it is argued that a website's "robots.txt" specifications binds web scrapers under some form of contract. This article shall attempt one such argument.

---

[53] This argument was attempted unsuccessfully in *PropertyGuru*, *supra* note 51. As a causal link between the alleged act of inducement and alleged breaches could not be established, the court did not see the need to examine the other elements of the tort. However, in the author's view, the tort of inducement of breach of contract would be an appropriate cause of action in a situation involving an application built upon web-scraped data, where the web scraping constituted a breach of a website's terms of service.

[54] See *Meltwater*, *supra* note 27 at para 49.

[55] Eliza Mik, "Contracts Governing the Use of Websites" (2016) Sing JLS 70 at 74-75.

[56] Kevin Conroy & John Shope, "Legal Analysis: Look Before You Click: The Enforceability of Website and Smartphone App Terms and Conditions" (2019) 63 Boston Bar J 23 at 23. Note also the categorical "click-wrap", "browse-wrap" or "sign-in wrap" approach from US case law, which Professor Mik criticises for failing to examine the intention of the contracting parties—see Mik, *ibid* at 73-74.

[57] See "Can a /robots.txt be used in a court of law?", *The Web Robots Pages* (blog), online: <http://www.robotstxt.org/faq/legal.html>.

From a first-principles perspective, it may be possible for the specifications of a "robots.txt" file to be incorporated as terms into the terms of service of a website,[58] due to the status of the REP as a widely-adopted industry standard—it might be reasoned that those who use web scrapers can be taken to subjectively know that they should comply with the REP.[59] However, as the REP is strictly voluntary in nature, this may militate against a finding that the parties had *intended* for the "robots.txt" specifications to form a part of the terms of service.[60]

V. PROHIBITIONS OF CONTRACTUAL OVERRIDES TO COPYRIGHT EXCEPTIONS

The above analysis suggests that, even though the use of web scrapers could be permitted by copyright exceptions under certain circumstances, contractual restrictions might ultimately impose the same (or even more stringent) restrictions upon the use of web scrapers. For example, a user of a web scraper might be able to rely on the TDM exception to escape liability for copyright infringement, but a website's terms of service might contain a restriction against web scraping, ultimately preventing the user from scraping the website. In this situation, the contractual restrictions imposed by the website owner have effectively overridden the copyright exceptions created by statute, rendering the copyright exceptions illusory (henceforth, these types of contractual restrictions shall be referred to as "contractual overrides").

In response to these contractual overrides, some jurisdictions have legislated for statutory prohibitions which render contractual overrides generally unenforceable (henceforth, these types of statutory provisions shall be referred to as the "prohibitions of contractual overrides"). Where these prohibitions of contractual overrides apply to the TDM exceptions, as well as the "fair use" and "fair dealing" exceptions, they shall be discussed below.

A. Contractual overrides to the text and data mining exceptions

[58] It is assumed that that the "robots.txt" file cannot constitute a contract in itself, either due to lack of consideration or lack of intention to create legal relations.

[59] See *Thornton v Shoe Lane Parking Ltd* [1971] 1 All ER 686 (a party is bound by terms which they either know of, or reasonable steps have been taken to give them notice of the terms; see also the famous "red hand rule", on the same page) at 690.

[60] *Carlill v Carbolic Smoke Ball Company* [1983] 1 QB 256 at 261-262.

It bears re-emphasis that the goal of the TDM exceptions is to remove the legal inhibitions created by copyright law in respect of TDM activities.[61] In that regard, allowing contractual overrides of the exceptions effectively leaves those who engage in TDM activities at square one.[62]

The UK and Singapore have responded by introducing blanket prohibitions of contractual overrides to their TDM exceptions.[63] In these two jurisdictions, the prohibitions of contractual overrides operate straightforwardly: they essentially provide that, to the extent that any contractual term is inconsistent with their respective TDM exceptions, the contractual term is rendered unenforceable.

Additionally, the EU provides an illustrative example of how a nuanced balance might be struck based on the relevant policy considerations, namely the promotion of innovation and ensuring freedom of contract: Under the *EU CDSM Directive*, the general TDM exception under Article 4 can be contractually overridden, but the specific exception under Article 3, which applies to TDM activities "conducted by research organisations and cultural heritage organisations" or "for the purposes of scientific research", cannot be.[64] This represents the policy that TDM activities can "in particular, benefit the research community and, in doing so, support innovation",[65] while for other organisations and purposes, "[r]ightsholders should remain able to license the uses of their works".[66]

B. Contractual overrides to the "fair use" and "fair dealing" exceptions

[61] See Part II.B of this paper.

[62] *Singapore Copyright Review Report*, *supra* note 31 at para 2.14.8. Note that this may be limited to particular situations where contractual overrides are likely to be prevalent—see e.g. *TDM – Legal Aspects*, *supra* note 33 at 13 (which notes that "[r]esearch and database providers often contractually override exceptions and limitations").

[63] *UK CDPA*, *supra* note 34, s 29A(5); *Singapore Copyright Act*, *supra* note 34, s 187(1)(c).

[64] *EU CDSM Directive*, *supra* note 34, articles 3 and 4 read with article 7(1).

[65] *Ibid*, recital 8.

[66] *Ibid*, recital 18.

While a small number of jurisdictions feature blanket prohibitions in their copyright statutes against contractual overrides to most (or all) of their copyright exceptions,[67] specific prohibitions of contractual overrides to the "fair use" and "fair dealing" exceptions are rare.[68]

Jurisdictions such as Ireland adopt straightforward blanket prohibitions which apply to their "fair use" and "fair dealing" exceptions.[69] The prohibition of contractual overrides found in Singapore's recently amended *Copyright Act* is also notable, as it provides for a robust set of requirements for contractual overrides to any copyright exception to be enforceable.[70] Of particular interest to users of web scrapers is the requirement that the contract must be individually negotiated before any contractual override may be enforceable; this requirement is unlikely to be made out for most contracts between the average users of web scrapers and website owners.

C. Contractual restrictions on the use of non-copyrighted code?

At this juncture, a technical loophole remains under the copyright statutes which contain prohibitions to contractual overrides: some contractual restrictions might apply to both copyrighted *and* non-copyrighted content hosted on a website,[71] while the prohibitions of contractual overrides, on their face, only apply to copyrighted content.

For users of web scrapers, this potentially means that contractual liability could result from the web scraping of HTML code that does *not* satisfy the requisite threshold for originality.[72] In contrast, the web scraping of sufficiently-original HTML code would not result in any liability in copyright or contract. This state of affairs is plainly unsatisfactory, and this article suggests one workaround to it through the TDM exceptions.

---

[67] See the examples cited in "Protecting Exceptions Against Contract Override: A Review of Provisions for Libraries" (27 November 2019), *IFLA*, online: <https://www.ifla.org/publications/node/92678> at 3.

[68] The author could not find any specific prohibitions of contractual overrides of the "fair use" and "fair dealing" provisions.

[69] *Copyright and Related Rights Act, 2000* (No. 28 of 2000, Ireland), s 2(10).

[70] *Singapore Copyright Act*, *supra* note 34, s 186(2)(a).

[71] See e.g. the clauses extracted at n 50 and n 51 of this article.

[72] It bears re-emphasis that the issue of whether users of websites are contractually bound by these terms is controversial—see generally Part IV of this paper. At the very least, this is a source of legal uncertainty for users of web scrapers.

It bears recalling that web scraping is a common precursor to TDM activities. Under such circumstances, the policy of the TDM exceptions, that TDM activities are beneficial to societies and economies,[73] applies equally to the mining of both copyrighted and non-copyrighted content. In this regard, the EU had considered clarifying that its TDM exception and corresponding prohibition of contractual overrides would be available for TDM activities concerning non-copyrighted content as well.[74] Therefore, it might be possible for the TDM exceptions to be interpreted by the courts to cover TDM activities concerning non-copyrighted content, though admittedly it would be awkward to interpret provisions in copyright statutes to cover uses of non-copyrighted material.

It would be much harder to extend the same argument beyond the TDM exceptions. For example, the "fair use" and "fair dealing" exceptions are ultimately premised on the general policy of copyright law,[75] and it would be conceptually unruly to extend prohibitions of contractual overrides of these exceptions over the web scraping of non-copyrighted content. In these situations, it might be possible to interpret the TDM exceptions even more broadly to encompass them, as the exceptions are non-exhaustively defined to begin with.[76]

VI. CONCLUSION: WHAT THIS MEANS FOR USERS OF WEB SCRAPERS

Having considered the two copyright exceptions which are most relevant to web scraping, the potential contractual restrictions which could bind users of web scrapers, and how the prohibitions of contractual overrides to the copyright exceptions operate—what does it all mean for users of web scrapers?

As a starting point, the answer depends on which jurisdiction's legal framework is at play. And it must be re-emphasised that the two legal controls discussed in this article—copyrights and contracts with websites—are not the only ones which exist in the web scraper's world.[77] Nonetheless, in jurisdictions which provide users of web scrapers with *both* copyright exceptions *and* prohibitions of contractual overrides, it is submitted that, once the requirements for the

[73] See Part III.B.1 of this article.

[74] *TDM – Legal Aspects*, *supra* note 33 at 6.

[75] See Part III.C of this article.

[76] See Part III.B.1 of this article.

[77] See Part I of this article.

relevant exceptions are met, web scrapers can be used without too much anxiety over liability for copyright infringement or breach of contract.

Of course, regardless of which legal frameworks are in place, web scrapers should always be used ethically. This exhortation is particularly important in the present age, where anyone with a computer and an Internet connection has easy access to the tools and tutorials needed to build their own web scrapers. Technological restrictions must be respected, "robots.txt" files should be followed, and when in doubt, it is always best to seek consent.[78] Through this article, it is hoped that those who wish to use web scrapers may do so with greater confidence in the legal frameworks underlying their activities—but without ever compromising our basic human decency.

[78] When seeking consent from website owners, they might offer access to their application programming interfaces ("APIs"), which could operate more efficiently than web scrapers!